8th International Workshop on Climate Informatics
September 19-21, 2018
Hosted by the National Center for Atmospheric Research in Boulder, CO

THE 2018 CLIMATE INFORMATICS HACKATHON: HURRICANE INTENSITY FORECAST

Sophie Giffard-Roisin[1], David Gagne[2], Alexandre Boucaud[3], Balázs Kégl[3], Mo Yang[3], Guillaume Charpiat[4] and Claire Monteleoni[1]

***Abstract*—The 2018 Climate Informatics hackathon focused on forecasting the hurricane intensities, and there was 35 participants. Specifically, the goal was to predict the intensity of tropical and extra-tropical storms (24h forecast) using information from past storms since 1979. The contest, dataset, methods, participants, successes, and challenges are discussed. The codes made to run the hackathon are freely available on the RAMP studio platform: www.github.com/ramp-kits/storm_forecast/.**

I. INTRODUCTION

Context: Cyclones, hurricanes or typhoons are words designating the same phenomena: a rare and complex event characterized by strong winds surrounding a low pressure area. Their trajectory and intensity forecasts are crucial for the protection of people and property. However, their evolution depends on many factors at different scales, altitudes and times, which leads to difficulties in their modelling. Today, the forecasts (track and intensity) are provided by a numerous number of guidance models[1]. Dynamical models solve the physical equations governing motions in the atmosphere. Statistical models, in contrast, are based on historical relationships between storm behavior and various other parameters. However, the lack of improvement in intensity forecasting is attributed to the complexity of tropical systems and an incomplete understanding of factors that affect their development. What is mainly still hard to predict is the rapid intensification of hurricanes: in 1992, Andrew went from tropical depression to a category 5 hurricane in 24h. Machine learning (and deep learning) methods have been only scarcely tested, and there is hope in that it can improve storm forecasts.

Corresponding author: S. Giffard-Roisin, sophie.giffard-roisin@mines-saint-etienne.org. [1] University of Colorado, Boulder USA. [2] National Center for Atmospheric Research, Boulder, Colorado. [3] Linear Accelerator Laboratory, Université Paris-Sud, CNRS. [4] Inria Saclay–Ile-de-France, LRI, Université Paris-Sud.

[1]NHC track and intensity models, www.nhc.noaa.gov/modelsummary.shtml, Accessed: 2018-07-04.

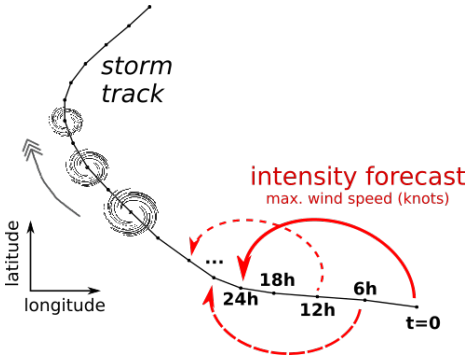

Fig. 1: Goal: estimate the 24h-forecast intensity of all storms.

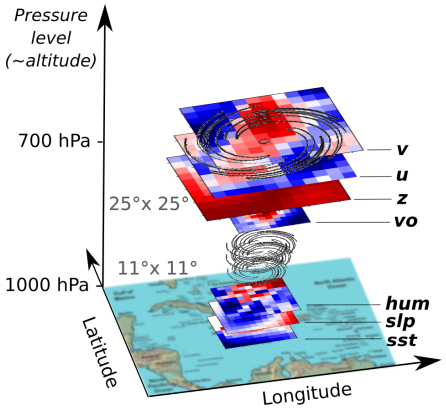

Fig. 2: Feature data: centered reanalysis maps of wind, altitude, sst, slp, humidity, vorticity.

Goal: This challenge proposed to design the best algorithm to predict for a large number of storms the 24h-forecast intensity every 6 hours. The (real) database is composed of more than 3000 extra-tropical and tropical storm tracks, and it also provides the intensity and some local physical information at each timestep [1]. Moreover, we also provide some 700-hPa and 1000-hPa feature maps of the neighborhood of the storm (from ERA-interim reanalysis database [2]), that can be viewed as images centered on the current storm location (see Fig. 2). The goal is to provide for each time step

8th International Workshop on Climate Informatics
September 19-21, 2018
Hosted by the National Center for Atmospheric Research in Boulder, CO

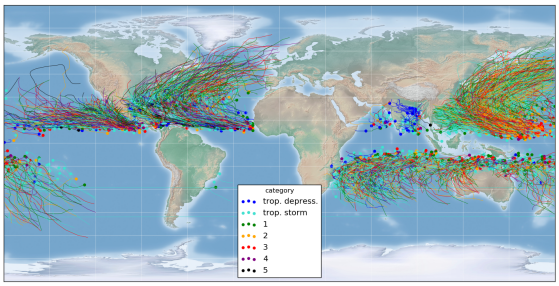

Fig. 3: Tracking database: more than 3000 tropical/extra-tropical storm tracks since 1979. Dots = initial position, colors = maximal strength (Saffir-Simpson scale)

of each storm (total number of time instants = 90,000), the predicted 24h-forecast intensity, so 4 time steps in the future (see Fig. 1).

II. DATA AND PIPELINE

Tracks and intensity: The raw storm track data used in this study is composed of more than 3000 extra-tropical and tropical storm tracks since 1979 extracted from the NOAA database IBTrACS [1], see Fig. 3. The tracks are defined by the 6-hourly center locations (latitude and longitude). They come from both hemispheres and the number of records per storm varies from 2 to 120 time steps. In total, the database counts more than 90,000 time steps. The intensity can be measured as the maximum sustained wind over a period of one minute at 10 meters height. This speed, calculated every 6 hours, is usually explained in knots ($1kt = 0.514m/s$) and is used to define the hurricane category from the Saffir-Simpson scale.

Features: The data consist of a set of features relative to one time step of every storm. First, the 0D features are a set of simple features : latitude, longitude, current windspeed, hemisphere, Jday predictor (Gaussian function of Julian day of storm init - peak day of the hurricane season), initial windspeed of the storm, last 12h windspeed change, storm basin, current distance to the land. Second, the reanalysis data (from the ERA-interim database [2]): at each time step, we extracted 7 grids (11x11 pixels) of meteorological parameters centered on the current storm location. Their choice is based on the forecast literature, on personal experience and on known hypotheses of storm strengthening. We provided 3 maps of 25 x 25 degrees (lat/long) at 700hPa-level pressure: the altitude z, and the u and v wind fields. These grids are subsampled to 11x11 pixels. We provide some more localized maps of 11 x 11 degrees at the

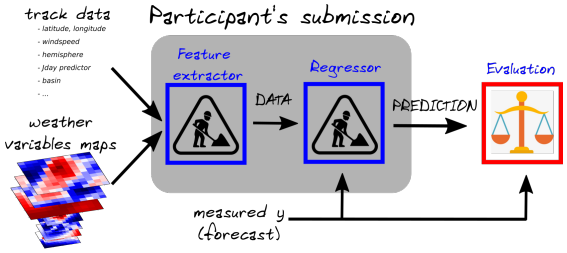

Fig. 4: Pipeline of the challenge.

surface: the sea surface temperature sst, surface level pressure slp, the relative humidity hum at 1000hPa (near surface). These grids are sampled to 11x11 pixels (1 pixel = 1 degree). We also provide the vorticity at 700hPa $vo700$, see Fig. 2.

RAMP platform: The Hackathon data, starting kit, and leaderboard were hosted through the Rapid Analytics and Model Prototyping (RAMP) website, developed by the Paris-Saclay Center of Data Science [3]. Unlike the machine learning contest site Kaggle, in which contest participants submit predictions from locally trained models, the RAMP site requires participants to submit Python code describing their feature extraction and machine learning model processes. The submitted code is then run on the RAMP web server to train and evaluate each participant's machine learning model. The source code for each team's submissions are then made visible to all participants so other teams can copy and modify it. This system encourages participants to build models that can run in a reasonable amount of time, discourages cheating, and increases the amount of cooperation among different teams.

Pipeline: Specifically, participants were asked to write two classes; a *feature extractor* and a *regressor* (see Figure 4). The 3000 storms have been randomly separated in a train set (half of the storms), a test set and a local starting kit. The local starting kit was available for download and includes only 1/4 storms of the total database. Once a submission was submitted to the platform, a cross-validation was performed on the train set and the result was shown on the *public leaerboard*. The test set (1/4 of the whole dataset) results were hidden, and the *private leaderboard* was revealed only at the end of the hackathon in order to prevent overfitting. The metric used is the RMSE (root mean square error) in knots across all storm time instants. We also made visible three other metrics: the mean absolute error (MAE, in knots), the MAE using only time instants corresponding to hurricanes (windspeed superior to 64 knots), and the relative RMSE on hurricanes. These

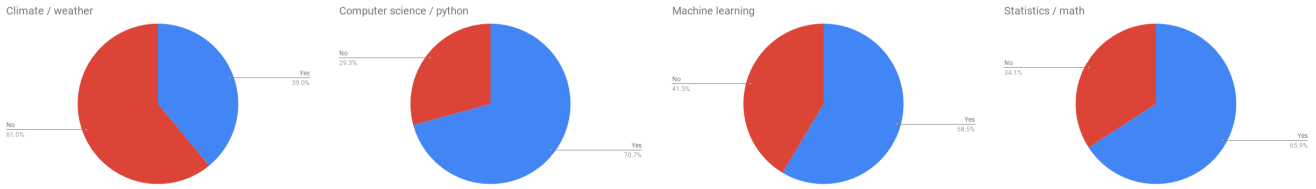

Fig. 5: Expertise of the 38 participants: (a) Climate / weather, (b) Computer science / python, (c) Machine learning, (d) Statistics / math.

metrics are interesting because the current forecasting practice is to exclude all other stages of development (e.g., extratropical, tropical wave...).[2]

III. PARTICIPANTS

There were 38 participants from which 18 were students and only one participant attended to a previous CI hackathon. In terms of expertise, less than 40% answered *Climate/weather*; while more than 70% answered *Computer science/python*. Based on their self-reported expertise, the participants were divided into 12 teams of 3 to 4 people so that each area of expertise would be covered by at least one team member.

IV. METHODS USED BY THE PARTICIPANTS

Feature extraction: Several ideas were developed by participants. A large majority did a feature selection or feature analysis (using significant variables or PCA). They were also many teams using historical features (from previous time steps of the same storm). If the 0D data was easy to use, the reanalysis data (small images) was subject to a variety of processing. If some methods used all the 11x11x7 reanalysis features independently, the majority pre-processed them according to climate science knowledge: for each map, they selected a subset of the mean, the maximum, the minimum, the central value or the min-max distance. Finally, some methods treated them as images (for convolutional neural networks).

Regression: A large majority used random forest regression with different hyperparameters, which seemed to perform well if interesting features were selected and if historical features were used. Some teams tried also gradient boosting or multi-layer perceptron regressor. Finally, few teams spent time in writing/training convolutional neural networks for the reanalysis maps, and a majority of them did not incorporate the 0D data in the features.

[2]More info at: www.nhc.noaa.gov/verification/verify5.shtml?

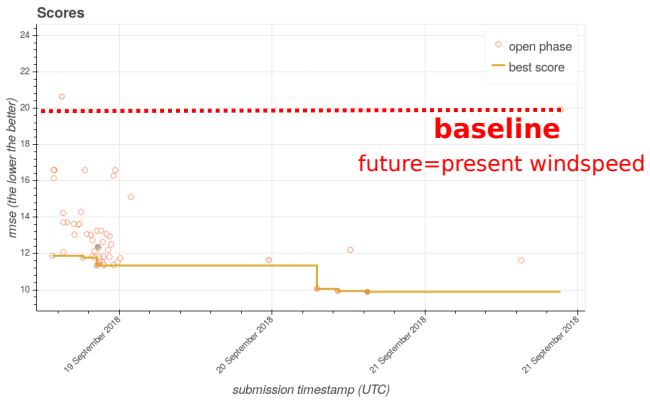

Fig. 6: Scores (RMSE) of the submissions wrt. time on the *public leaderboard* (cross-validation score using the online training set, score visible during the event).

V. RESULTS

Public leaderboard results: We show in Figure 6 the scores of the *public leaderboard* (cross-validation score using the training set, score visible during the event) with respect to time. There were more than 50 online submissions, mostly concentrated during the hackathon day, even if some participants continued to submit during the following 2-day workshop. We can see that the large majority of the submissions are beating the baseline which RMSE = 20 knots (the baseline consists in using the current windspeed as prediction for the 24h-forecasted windspeed). During the first day (hackathon day), the best results were found when using random forests methods. We can see a breakthrough in the lowest RMSE (yellow line) after the Sept. 20 due to the tuning of convolutional neural networks by few teams. The final best result on the public leaderboard was: RMSE = 9.9 knots corresponding to a mean absolute error MAE = 6.3 knots.

Private leaderboard results: At the end of the open phase, the best submission of every participant on the *public leaderboard* is used to predict the results on

8th International Workshop on Climate Informatics
September 19-21, 2018
Hosted by the National Center for Atmospheric Research in Boulder, CO

Rank	Team	Submission name	Rank move	RMSE	MAE (knots)	train time	test time	reg. method
1	yeliz	team8_GBR_map	+5	13.6	9.7	31 s	5 s	grad. boosting
2	pjn	test	+2	13.6	9.8	3099 s	20 s	random forest
3	changgt	divergence_rf_fixed	+2	13.8	10	39 s	13 s	random forest
4	ftonini	team8_RF_maps_sc_num	+6	13.9	10	93 s	9 s	random forest
5	djgagne	keras_big_cnn_djg	+8	13.9	10.1	164 s	11 s	conv. nets
21	-	**baseline**	-	20.9	14.3	1 s	4 s	*baseline* *

TABLE I: *Private competition leaderboard*: winning first 5 teams on the hidden test set. *Rank move* = rank difference of the submission with respect to the *public leaderboard*. * The baseline consists in using the current windspeed as prediction for the 24h-forecasted windspeed.

the *hidden* test set, leading to the *private competition leaderboard* shown in Table I. The best RMSE (13.6 knots) is obtained with a gradient boosting method, and they extracted all 0D features together with the mean and the center of every reanalysis map. The second best uses random forests using historical 0D data. We also show the mean absolute error, where the best team reaches 9.7 knots. For comparison, we recall that a storm is considered a hurricane if its max. windspeed exceeds 65 knots. We can see that the baseline on this new test set has comparable score (RMSE = 20.9) to the one on the public leaderboard (RMSE = 19.9). Finally, the column *Rank move* tells us the rank difference of the submission with respect to the *public leaderboard*, accessible during the hackathon.

VI. DISCUSSION

Overfitting: We can see from the *Rank move* of Table I that the winning teams were not the ones having the best results on the *public leaderboard*. Moreover, the best score of the test set (RMSE = 13.6) is very different from the one coming from cross-validation on the online training set (RMSE = 9.9, see Figure 6). Even if the online training set was not directly accessible (only the submissions scores and the codes were), this seem to indicate an overfitting of the training set by many teams. Each team was able to submit every 15 minutes. This type of task and dataset, having a limited number of examples but with a large variability, is known to be more subject to overfitting. We can also note that different types of methods are in the top 5 (gradient boosting, random forest and convolutional neural nets) which seem to indicate that overfitting is not restricted to one method. We can clearly see here the importance of having a hidden test set for such applied research problems.

Collaborative research: Nonetheless, this hackathon was a good example of collaborative research on an important climate topic, and the results are very promising. Teams did benefit from the expertise of their members and all said to have learned a great deal. Moreover, the nature of the RAMP workflow enabled participants to to re-use each other's ideas and codes. We decided in this hackathon that the submission codes of other teams would be visible from the beginning. However, we think that this might lead to even more overfitting, as teams looked at the best current submission and tuned it. Thus, we can see some benefit in opening the visibility only after some time in order to have more diverse submissions.

Dissemination: We finally want to point out that the problem set can be re-used, by asking the RAMP team to open a new event (visit www.ramp.studio). The codes made to use this RAMP are also freely available on github (www.github.com/ramp-kits/storm_forecast/). Finally, all the data was freely available online before processing it from the IbTrACS and the ERA-interim databases [1], [2]. The processed data can also be shared to anyone by sending an email to the corresponding author.

REFERENCES

[1] K. R. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, "The international best track archive for climate stewardship (ibtracs) unifying tropical cyclone data," *Bulletin of the American Meteorological Society*, vol. 91, no. 3, pp. 363–376, 2010.

[2] D. P. Dee, S. M. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, d. P. Bauer, *et al.*, "The era-interim reanalysis: Configuration and performance of the data assimilation system," *Quarterly Journal of the royal meteorological society*, vol. 137, no. 656, pp. 553–597, 2011.

[3] B. Kégl, A. Boucaud, M. Cherti, A. Kazakci, A. Gramfort, G. Lemaitre, J. Van den Bossche, D. Benbouzid, and C. Marini, "The ramp framework: from reproducibility to transparency in the design and optimization of scientific workflows," 2018.